A concurrent curve strategy for formant tracking

Yves Laprie

A concurrent curve strategy for formant tracking

Yves Laprie

LORIA CNRS
Nancy, France
Yves.Laprie@loria.fr

Abstract

Although automatic formant tracking has a wide range of potential applications it is still an open problem. We previously proposed the use of active curves that deform under the influence of the spectrogram energy. Each formant was tracked independently and a complex strategy was required to guarantee the overall formant tracking consistency. This paper describes how the interdependency between formants can be incorporated directly during the deformations of formant tracks. Iterative processes attached to each formant are interlaced. We experimented two strategies. The first consists in partitioning the spectrogram into exclusive regions, each region affiliated to a given formant. The second consists in adding a repulsion force between formants that prevent formant tracks to merge together. It turns out that the second strategy is more robust and does not necessitate a complex control strategy.

1. Introduction

As formants directly derive from the geometrical shape of the vocal tract, they may be exploited to identify sounds pronounced, especially vowels and other vocalic sounds. Formant tracks are utilized to pilot formant synthesizers [5], to study coarticulation effects, vowel perception, articulatory phenomena and in some rare cases to provide a speech recognition with additional data [3].

Given the potential interest of formant data numerous works have been dedicated to the design of automatic formant tracking algorithms. Nature and complexity of the problem explains the success of dynamic programming algorithms [9, 10]. The first steps of these algorithms is the extraction of formant candidates at each frame of the speech signal. The second stage is dynamic programming that utilizes the evaluation of transition costs between two frames. Other algorithms aim at explaining the acoustic signal [1] or the spectrogram energy. In [7] we showed how active curves could be used to track formants. The underlying idea is to deform initial rough estimates of formants under the influence of the spectrogram to get regular formants tracks close to lines of spectral maxima which are potential formant tracks. A formant track is thus represented by a curve $t : [t_i, t_f] \rightarrow \mathbb{R}^2, t \rightarrow (t, F(t))$ in the time frequency domain, t_i and t_f are times of the beginning and end of the formant track, and $F(t)$ is the frequency of the formant at time t. The compromise between proximity to spectral peaks and regularity is given by the following functional which has to be minimized

$$E(F) = \begin{array}{c} -\int_{t_i}^{t_f} E_{Spectro}(t, F(t))dt \\ +\lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \end{array} \quad (1)$$

where the overall energy $E(F)$ has to be minimized. The first term represents the spectrogram energy along the formant track. It is thus all the bigger since the curve is close to a line of spectral peaks. The second term represents the length and the curvature of the track and is thus all the smaller since the curve is regular. α influences the curve length, β its curvature and λ the compromise between the spectrogram energy explained by the formant tracks and the smoothness of the curve.

The main strength of this approach is that formant tracks are moving towards tracks of spectral peaks, i.e. formants, while guaranteeing that formants tracks are sufficiently smooth and close to spectral peaks.

To minimize $E(F)$ Eq. 1 is derived with respect to F by using Euler equations which gives

$$-\alpha F^{(2)} + \beta F^{(4)} - \frac{1}{\lambda}\frac{\partial E_{Spectro}}{\partial F} = 0 \quad (2)$$

Finite difference approximation of the derivatives leads to an equation, which can be written in the matrix form

$$AF = \frac{1}{\lambda}\frac{\partial E_{Spectro}}{\partial F} \quad (3)$$

where matrix A corresponds the smoothness term (see [7] for further details). However, matrix A is ill conditioned and it is therefore necessary to use an iterative numerical process which builds a sequence $F^n = [F_0^n, F_1^n ... F_N^n]$ of curves that converges to the solution of Eq. 1. This process depends on a parameter γ. Eq. 1 is therefore solved iteratively by using

$$F^n = (A+\gamma I)^{-1}(\gamma F^{n-1} + \frac{1}{\lambda}\frac{\partial E_{Spectro}}{\partial \mathcal{F}}(t, F^{n-1}(t))) \quad (4)$$

Each formant curve deforms under the influence of the spectrogram independently of the other formant curves, what requires a complex control strategy to manage interactions between formants. The main difficulty is when two formants are competing with each other to catch the energy of one spectral peak. This problem occurs when one spectral peak is too weak compared to the other and leads the two formants tracks to get closer to the prominent spectral peak. Another difficulty is the initialization of the tracks that requires the construction and the labelling of elementary tracks.

We report here how the deformation equation can be modified to incorporate interdependency between formants. This way, formants are deforming by taking into account the deformations of their neighboring formants with two advantages: a better coverage of the spectrogram energy, a simpler and more robust control strategy. Moreover, the initialization stage can be substantially simplified because the interdependency of formant tracks enables an more dynamic exploration of solutions than that possible with the labelling of elementary tracks based on a static strategy.

We present two methods for adding the interdependency between formant tracks. The first consists in partitioning the

spectrogram into exclusive domains, one for each formant track. The second consists in adding a repulsion term between formant tracks in Eq. 1. The spectral analysis used to compute spectrograms plays a central role since it defines the energy field where formant tracks deform. We thus report how a true envelope algorithm derived from cepstrum improves formant tracking.

2. Interdependency of formant tracks

Basically, the interdependency is achieved by interlacing the iterative processes of formant tracks. Instead of conducting the iterative processes attached to formants one after the other there is one general iterative process. At each general iteration, all the formants F_i are examined, one after the other, by carrying out one iteration that influences neighboring formants tracks F_{i-1} and F_{i+1}.

2.1. Spectrogram partition strategy

The principle is to build a partition of the spectrogram into exclusive regions, one for each formant. Each formant deformation is thus controlled by a specific region and regions are modified according to the deformation of formants tracks. The general layout of the algorithm is as follows:

1. initialization of the M formant tracks. The i^{th} formant track is the moving average of the i^{th} LPC root.
2. compute one iteration of the deformation for each of the M formant tracks and modify the partition of the spectrogram accordingly
3. go to step 2 until the formant tracks stabilize.

Sun [8] proposed a similar method applied to cubic splines. The partition was achieved through the calculation of the probability of a spectral frequency to belong to a given formant.

Preliminary experiments showed that the probability could give rise to very narrow regions (i.e. with a very small standard deviation) for high energy formants, and consequently, to a non-relevant partition of the spectrogram. Therefore, we accepted a non probabilistic frequency-to-formant affiliation function $a_i(t,z)$ given the affiliation of the frequency z at time t to the formant F_i:

$$a_i(t,z) = exp(c \times ln(4\frac{(z - F_{i-1}(t))(F_{i+1}(t) - z)}{(F_{i+1}(t) - F_{i-1}(t))^2}))$$

This function returns 1 at the center of the domain affiliated to the i^{th} formant and the constant c controls the decrease of the affiliation when the frequency moves away from that of the formant. Affiliation of first and last formant are slightly different so to obtain an affiliation scheme like that presented by Fig. 1.

Despite its simplicity this tracking strategy turns out to be as efficient as that reported in [7]. However, we noticed that the partition becomes non-relevant when a formant does not correspond to a peak in the spectrum. This often happens in the case of a nasalized sound. Fig. 2 shows the example of the sequence /yRã/. The line of spectral peaks corresponding to F1 is strongly dominated by that of F2 near 1000 Hz that attracts the F1 track. By moving towards higher frequencies F1 track thus pushes away the F2 track and explains the erroneous result of Fig. 2a. The weak point here is that the F2 track has reached a correct position but cannot push back the move of F1 track.

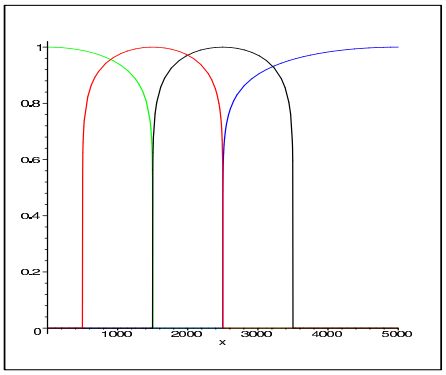

Figure 1: Affiliation profile with F1=500 Hz, F2=1500 Hz, F3=2500 Hz et F4=3500 Hz

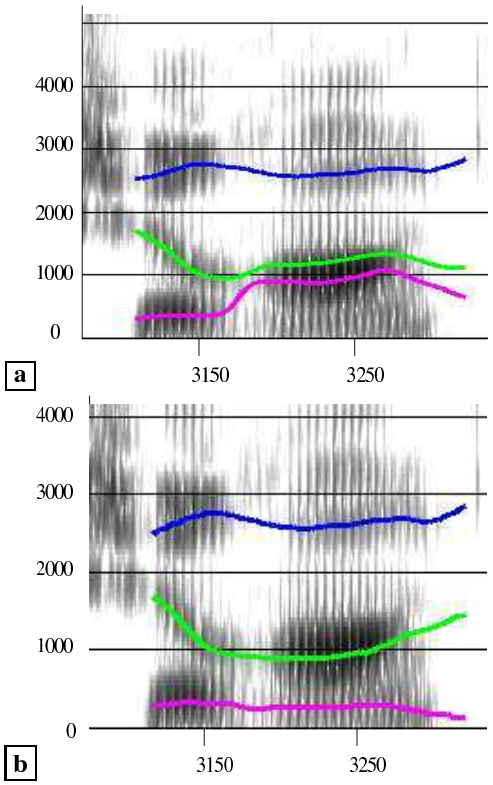

Figure 2: Formants tracks obtained by the partition strategy (**a**) by the repulsion strategy (**b**)

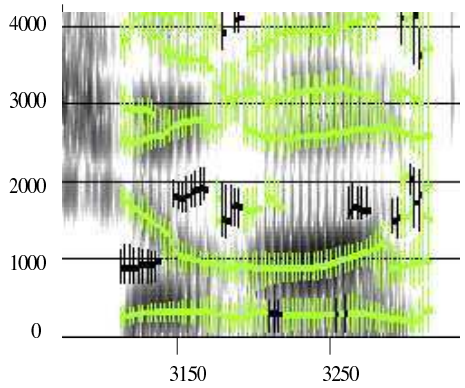

Figure 3: Peaks and bandwidth obtained by the true envelope algorithm. Black peaks correspond to very small peaks.

2.2. Repulsive tracks strategy

We thus designed a new strategy that incorporates repulsion forces to keep formant tracks away form each other. This is achieved by incorporating a repulsive term in Eq. 1 which becomes

$$\begin{aligned} E(F) = \quad & -\int_{t_i}^{t_f} E_{Spectro}(t, F(t)) \\ & +\mu \sum_n E_{Spectro}(t, F_n(t)) \times exp^{-(\frac{F_n(t)-F(t)}{s_n})^2} dt \\ & +\lambda \int_{t_i}^{t_f} \alpha |F'(t)|^2 + \beta |F''(t)|^2 dt \end{aligned} \tag{5}$$

where μ is the weight of the repulsive force and n gives the neighboring tracks. For the second formant F_2 n corresponds to F_1 and F_2 and for F_3 n corresponds to F_2. For the first formant F_1 n corresponds to F_2 and an artificial fixed formant (at 0 Hz) to prevent F_1 from becoming negative. The constant s controls the scope of the repulsion and has been set to 300 Hz except for the artificial zero formant for which it has been set to 100 Hz. Note that the repulsion is proportional to the energy of the formant that exerts repulsion. We accepted $\mu \times exp^{-(\frac{F_n(t)-F(t)}{s_n})^2}$ as the repulsion term rather than one of the form of $\mu \frac{1}{(F_n(t)-F(t))^2}$ because the scope and the shape of the former can be more easily adjusted and does not give rise to too strong values.

By applying the Euler equations to the new form of the functional $E(F)$ we obtain the matrix equation that corresponds to Eq. 3

$$\begin{aligned} AF = \quad & \frac{1}{\lambda} \frac{\partial E_{Spectro}}{\partial F} \\ & +2\frac{\mu}{\lambda} \sum_n E_{Spectro}(F_n) \times (F - F_n) \times exp^{-(\frac{F_n - F}{s_n})^2} \end{aligned} \tag{6}$$

As shown by Fig. 2b this strategy corrects the weakness of the spectrogram partition strategy.

3. Spectrogram calculation

The spectral analysis used to compute spectrograms plays a central role since formant tracks deform under the influence of the spectrogram. It thus has to render formants faithfully. Therefore, linear prediction cannot be used, at least for the deformation step, because it does not fit spectra of nasalized sounds correctly. Cepstral smoothing behaves better but d not offer a sufficient frequency resolution. One of the reasons is that the behavior of the cepstral analysis slightly depends on the location of the analyzing window with respect to pitch periods, what results in the instability of weak spectral peaks in terms of frequency and energy. The true envelope algorithm proposed by Imai and Abe [6] (also described by Halle [4]) solves this problem by providing a spectrum that approximates spectral peaks, i.e. harmonics in the case of voiced speech.

Let S be the narrow band spectrum,
$V^{(1)} = \hat{S}$ ($\hat{S}$ is the cepstrally smoothed spectrum),
$E^{(1)} = g(S - \hat{S})$ where $g(y)$ = if $y > 0$ then y else 0
$E^{(1)}$ represents the positive difference of S with respect to $\hat{S}$
$\hat{E}^{(1)}$ is the cepstral smoothing of this difference (see Fig. 4) which is added to $\hat{S}$ so as to put the smoothed spectrum closer to peaks. The algorithm works as follows (DFT is the discrete Fourier transform and IDFT the inverse transform)

1. initial solution
   $\hat{E}^{(1)}(k) = \sum_{m=0}^{N-1} e_m^{(1)} h_m cos(\frac{2}{N} mk)$ where N is the order of the Fourier transform, $e^{(1)} = IDFT(E^{(1)})$, h_m is the liftering window and k the frequency bin computed.
2. iteration $i+1$
   let $V^{(i)}$ the envelope obtained at the previous stage, $E^{(i)}$ and $\hat{E}^{(i)}$ the corresponding positive difference and smoothing of the difference.
   $V^{(i+1)} = V^{(i)} + \hat{E}^{(i)}$
   $E^{(i+1)} = g(E^{(i)} - (1+\alpha)\hat{E}^{(i)})$ where α is an acceleration coefficient.
   $\hat{E}^{(i+1)} = DFT(h(IDFT(E^{(i+1)})))$
3. goto step 2 or stop (in practice 6 iterations are sufficient).

True envelope and discrete cepstrum proposed by Gallas and Rodet [2] are very similar ; the advantage of the former method is to not require the knowledge of spectral peaks to be interpolated.

4. Determination of initial formant tracks

Although the two strategies presented above achieve an efficient interdependency between formants initial formant estimates have still to be provided. However, the overall result will not be as narrowly linked to the initialization as it was the case in our previous approach [7]. As explained in the previous section, we accepted the true envelope spectrogram for generating formant track deformations because it approximates spectral peaks with a better faithfulness than linear prediction. On the other hand, like cepstral smoothing, the true envelope may give less peaks than formants. We therefore chose to construct initial estimates of the i^{th} formant tracks by taking the moving average of the i^{th} LPC root. The average window is sufficiently long (250 ms) to attenuate effect of outliers and local errors and deformation forces created by the true envelope spectrogram compensate for the lack of faithfulness of the LPC as shown in Fig. 2.1 and 2.

5. Results

We manually examined results over 18 repetitions of a small story ("La bise et le soleil") uttered by French speakers. The phonetic transcription comprises 384 phonemes, 39 of them are nasal French vowels or nasalized sounds (/m,n/). Although the nasalized sounds only represent 10% of the total they make tracking very difficult even for a human expert because some of them are contiguous, “son manteau” /s õ m ã t o/, for instance. We found no gross error in non-nasalized sounds except

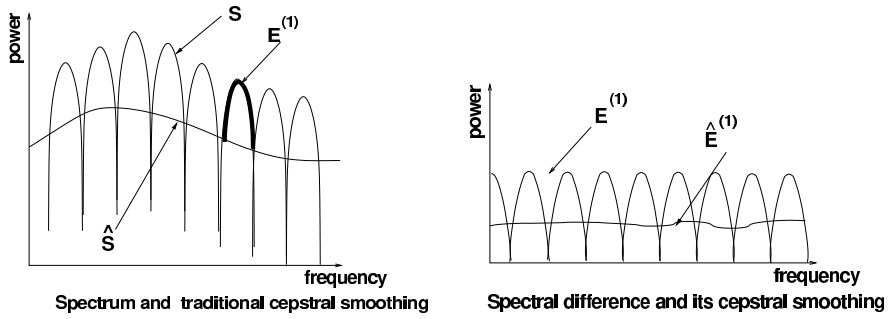

Figure 4: Principle of the true envelope computation

in two cases: **(i)** for F2 and F3 of an non-accentuated /y/ which are very close together and fairly weak, **(ii)** at the end of vocalic segments lengthened by sonorants, like /ɔR/ for instance. In the latter case the error concerns the slope of F1 or F2 at the very end of the vocalic segment. Tracking performs also well for nasal vowel /œ̃/, and /õ ã/ provided that these two vowels are not surrounded by other nasalized sounds or the nasalization is not too strong. On the other hand, when the nasalization is very marked one or two extra spectral peak lines appear below 3000 Hz results are often erroneous. Our corpus comprises 18×4 sequences /õ m ã/. It turns out that only one third of the corresponding formant patterns have been correctly detected. Most frequent errors due to the vanishing of one of the two formants below 1100 Hz and to a lesser extent to the existence of 5 spectral peaks below 3000 Hz. In the latter case the three formant tracks are unable to explain these peaks and tracks jumps from one formant to another according to the their corresponding peak amplitude.

6. Discussion and conclusion

The incorporation of interdependency between formant tracks in the numerical scheme turns out to be an efficient way for guaranteeing the overall consistency of the formant tracking. Compared to dynamic programming approaches, the advantage is to take into account the whole energy distribution of the spectrogram and not only peaks. However, as for other approaches of formant tracking, the success depends on the adequation between the number of formants tracked and the number of spectral lines. This is particularly sensitive for nasals and nasalized sounds which may present extra formants. In this case some formant may be missed. In order to detect these errors we will investigate how the ratio of energy integrated by formant tracks can be exploited to evaluate results. This evaluation will enable the segmentation of the signal and the optimal number of formant tracks to be determined.

7. References

[1] I. Bazzi, A. Acero, and L. Deng. An expectation maximization approach for formant tracking using a parameter-free non-linear predictor. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'2003, Hong-Kong*, May 2003.

[2] T. Gallas and X. Rodet. Generalized functionnal approximation for source-filter system modelling. In *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[3] Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden makov mdels for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, Seattle, USA, May 1998.

[4] P. Halle. Techniques cepstrales améliorées pour l'extraction d'enveloppe spectrale et la détection du pitch. In *Actes du séminaire "Traitement du signal de parole"*, pages 83–93, Paris, 1983.

[5] J. N. Holmes. Formant synthetisers: cascade or parallel ? *Speech Communication*, 2:251–273, 1983.

[6] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Trans. IECE*, J62-A(4):217–223, 1979 (in japanese).

[7] Y. Laprie and M.-O. Berger. Cooperation of regularization and speech heuristics to control automatic formant tracking. *Speech Communication*, 19(4):255–270, October 1996.

[8] Don X. Sun. Robust estimation of spectral center-of-gravity trajectories using mixture spline models. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, volume 1, pages 749–752, Madrid, Spain, September, 1995.

[9] D. Talkin. Speech formant trajectory estimation using dynamic programming with modulated transition costs. *Journal of the Acoustical Society of America*, S1:S55, March 1987.

[10] K. Xia and C. Epsy-Wilson. A New Strategy of Formant Tracking based on Dynamic Programming. In *International Conf. on Spoken Language Processing - ICSLP2000, Beijing, Chine*, October 2000.